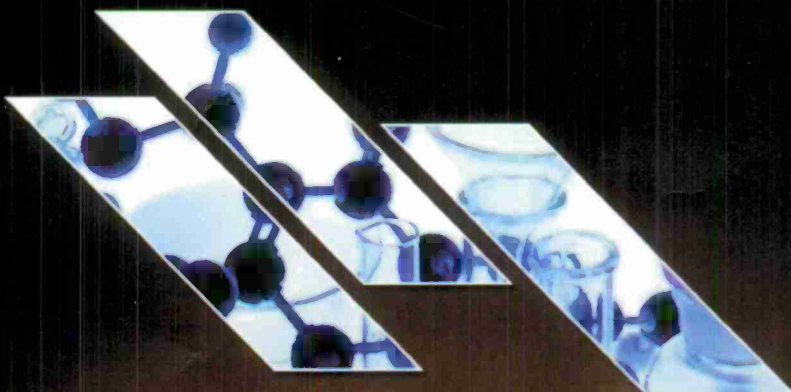
EDGEWOOD CHEMICAL BIOLOGICAL CENTER

U.S. ARMY RESEARCH, DEVELOPMENT AND ENGINEERING COMMAND

Aberdeen Proving Ground, MD 21010-5424

ECBC-TR-917

EVALUATION OF THE DEGREE OF SEPARATION BETWEEN TWO DATA POPULATIONS WITH STATISTICAL ALGORITHMS



Waleed M. Maswadeh
A. Peter Snyder

RESEARCH AND TECHNOLOGY DIRECTORATE

April 2012

Approved for public release; distribution is unlimited.

U.S. ARMY RDECOM

Disclaimer

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorizing documents.

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* XX-04-2012	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* Aug 2010 - Apr 2011
4. TITLE AND SUBTITLE Evaluation of the Degree of Separation between Two Data Populations with Statistical Algorithms		**5a. CONTRACT NUMBER** **5b. GRANT NUMBER** **5c. PROGRAM ELEMENT NUMBER**
6. AUTHOR(S) Maswadeh, Waleed M.; and Snyder, A. Peter		**5d. PROJECT NUMBER** **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) DIR, ECBC, ATTN: RDCB-DRD-P, APG, MD 21010-5424		**8. PERFORMING ORGANIZATION REPORT NUMBER** ECBC-TR-917
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)		**10. SPONSOR/MONITOR'S ACRONYM(S)** **11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT-LIMIT 200 WORDS
This report investigates various existing statistical methods available in the literature and new statistical methods for evaluating the degree of separation between two peaks or two populations of data. The algorithms evaluated in this study include the direct percentage of overlap between the two populations of data, the Kolmogorov-Smirnov (K-S) test, the area between the receiver operating characteristic (ROC) curve and diagonal line, and the ROC curve length (LROC). These algorithms are compared to the standard reference probability distribution for each data profile to be separated. Evaluations of the algorithms are presented in order to determine the relative degree of separation between the two peaks or distributions. The LROC is determined to provide the best estimation of the degree of separation compared to the other methods.

15. SUBJECT TERMS

Multivariate analysis	Univariate analysis	Fisher iris flower data set
Sepal	Petal	Receiver

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					Renu B. Rastogi
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U	UL	40	(410) 436-7545

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

20130117069

Blank

PREFACE

The work described in this report was started in August 2010 and completed in April 2011.

The use of either trade or manufacturers' names in this report does not constitute an official endorsement of any commercial products. This report may not be cited for purposes of advertisement.

This report has been approved for public release.

Blank

CONTENTS

1. INTRODUCTION 9

2. EXPERIMENTAL PROCEDURE 9

3. RESULTS AND DISCUSSION 10

3.1 Profiles of Different Data Distributions 10
3.2 Two Identical Square-Wave Profiles 11
3.3 Two Square-Wave Profiles with Different Peak Widths 13
3.4 Probability Functions as Reference Sources from Experimental Data 14
3.5 Application of Probability Functions on Square-Wave Profiles 15
3.6 Application of Probability Functions on Square-Wave Profiles with Different Peak Widths 18
3.7 Application of Probability Functions on Triangular Data Profiles 18
3.8 Application of Probability Functions on Two Triangular Profiles with Different Peak Widths 20
3.9 Application of Probability Functions on Skewed Profiles 21
3.10 Application of Probability Functions on Triangular Profiles with a Bimodal Distribution 23
3.11 Other Distribution Profiles 24

4. CONCLUSIONS 27

LITERATURE CITED 29

GLOSSARY 30

APPENDIXES

A. KOLMOGOROV-SMIRNOV TEST 31

B. THE AREA BETWEEN ROC CURVE AND THE DIAGONAL LINE ACD AND LENGTH OF THE ROC CURVE 33

FIGURES

1. Evaluation of the percent degree of separation between distributions A and B 10

2. Degrees of overlap for the A and B square-wave distributions of data having identical profiles 11

3. Relationships between the calculated percent average separation (using overlap area) vs. the K-S test (circles), ACD (small diamonds), and $LROC_O$ (big diamonds) values from Table 1 12

4. Degrees of overlap for two different data distribution profiles 13

5. True reference value for degree of separation with 100% confidence separation (a) for distribution A ($\mathbf{A_{100}}$) and (b) for distribution B ($\mathbf{B_{100}}$) (*x* axis) vs. the (a) percent separation using overlap area A/100 (open circles, *y* axis) and the 90% confidence $\mathbf{A_{90}}$/100 values (closed circles) and (b) the percent separation B/100 (closed circles) and 90% confidence values $\mathbf{B_{90}}$/100 (closed circles) (*y* axis), respectively 17

6. Relationships between $\mathbf{AB_{90}}$ vs. the K-S test (circles), ACD (small diamonds), and $LROC_O$ (large diamonds) values for Figure 4(a–d) 17

7. Degrees of overlap for two data distributions with the same triangular profile 19

8. Relationships between the $\mathbf{AB_{90}}$ vs. the K-S test (small diamonds), ACD (circles), and $LROC_O$ (large diamonds) values from the Table 4 data 20

9. Degrees of overlap for two different distributions of data 20

10. $\mathbf{AB_{90}}$ values plotted against the K-S test (circles), ACD (small diamonds), and $LROC_O$ (large diamonds) values from Table 5 21

11. Degrees of overlap for two PSDs 22

12. Relationships between the $\mathbf{AB_{90}}$ and the K-S test (circles), ACD (small diamonds), and $LROC_O$ (large diamonds) values in Table 6 23

13. Degrees of overlap for two different distributions of data where profile A is a triangular distribution and profile B is a bimodal distribution 23

14. Relationships between the $\mathbf{AB_{90}}$ vs. the K-S test (circles), ACD (small diamonds), and $LROC_O$ (large diamonds) values in Table 7 24

15. Plotted values from the K-S test, ACD, and $LROC_O$ analyses from Figure 1(a–i) at various overlap stages plotted against the $\mathbf{AB_{90}}$ values 25

16. Figure 14 results with the data from Figure 1(h,i) removed 26

TABLES

1. Overlap Analyses of the Distributions of Data in Figure 2(a–d) 12

2. Calculated Values of the Overlap Area, Percent Overlap, and Percent Separation Using the Overlap Area from Figure 2(a–d). 16

3. Overlap Analyses of the Two Distributions of Data in Figure 4(a–d) 18

4. Overlap Analyses of the Distributions of Data in Figure 6 19

5. Overlap Analyses of the Two Distributions of Data in Figure 9(a–d) for the $\mathbf{A_{90}}$, $\mathbf{B_{90}}$, $\mathbf{AB_{90}}$, K-S test, ACD, and $LROC_O$ Values 21

6. Overlap Analyses of Two Distributions of Data in Figure 11(a–d) 22

7. Overlap Analyses of the Distributions of Data in Figure 13(a–d) 24

Blank

EVALUATION OF THE DEGREE OF SEPARATION BETWEEN TWO DATA POPULATIONS WITH STATISTICAL ALGORITHMS

1. INTRODUCTION

There is a need in the scientific community for a method that can measure the degree of separation between two population distributions or two peaks. Such study and method development may help to improve many research and development areas as well as statistical analysis techniques such as multivariate data analysis. Some of the following tests, available in the literature, that are commonly used for the separation and discrimination of two peaks are

- The direct percentage of overlap between two populations (simple overlap area)
- The Kolmogorov-Smirnov Test (K-S test)
- The area under the curve (AUC) of the receiver operating characteristic (ROC) curve (Metz, 1978)

The K-S test has been extensively used by researchers (Wilcox, 1997) and it is presented in Appendix A. The ROC curve is utilized in the literature to a significant degree (Fawcett, 2006; Hanley and McNeil, 1982). Two additional methods are introduced in this report:

- The area under the ROC curve referenced to the diagonal line (ACD)
- The length of the ROC curve (LROC)

The ACD and LROC methods are discussed in detail in Appendix B. The overlap area and K-S test provide only a single value with respect to separation determinations. However, the ACD and LROC analyses provide not only a measurement for the degree of separation and discrimination, but they also yield additional quantitative statistical information such as sensitivity and selectivity.

2. EXPERIMENTAL PROCEDURE

Synthetic sets of various population profiles were calculated and displayed to explore the degree of separation between any two distributions (two peaks or two populations). Data were analyzed with various algorithms. The objective was to arrive at the best algorithm that can accurately estimate the degree of separation for two distributions. The mathematical approaches of the analyzed algorithms are presented in Appendices A and B.

3. RESULTS AND DISCUSSION

Various population profiles of different shapes were synthetically calculated and displayed to explore the degree of separation between two distributions (two peaks or populations). Figure 1(a–i) shows various distribution profiles constructed to evaluate the degree or percent of separation between two distributions (i.e., distributions A and B in Figure 1[a–i]).

3.1 Profiles of Different Data Distributions.

The simplest distribution profile is a square wave (continuous uniform or rectangular distribution). Two such peaks are shown in Figure 1a with the same full width at half height (FWHH), and two peaks with different FWHHs are presented in Figure 1b. The next simplest distribution profile is triangular, and two such profiles are shown in Figure 1c with the same FWHH. Two triangular-distribution profiles with different FWHH are presented in Figure 1d. A triangular-distribution profile approximates a binomial-distribution profile.

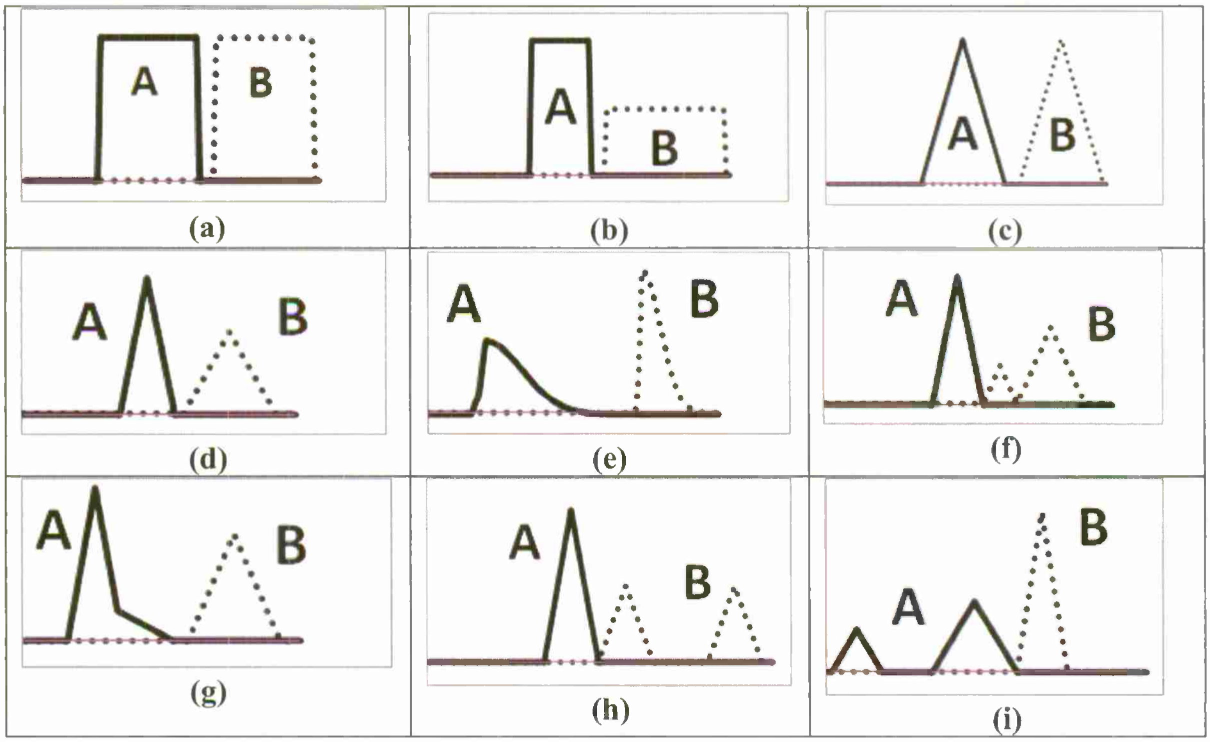


Figure 1(a–i). Evaluation of the percent degree of separation between distributions A and B.

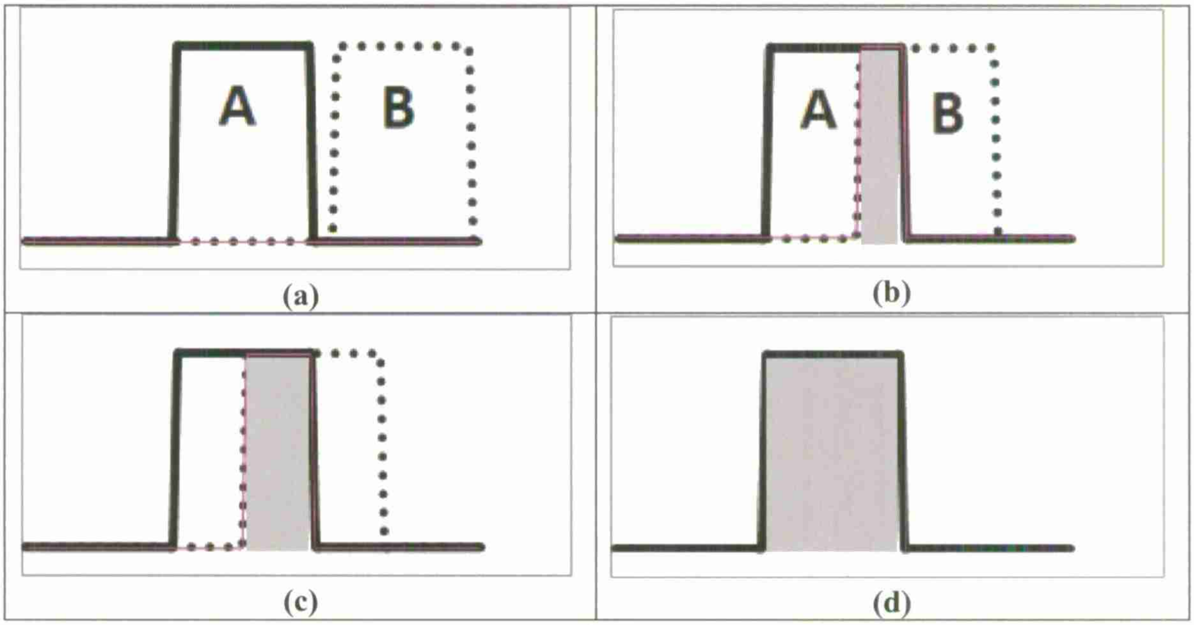
The main objective in this study was to find a method that is best suited for measuring the percent degree of separation between two data sets regardless of the nature of the distribution profile shape (rectangular, Gaussian, Skew normal, Binomial, Bimodal, Negative binomial, chi, or Poisson). Figure 1(a–i) shows data distributions labeled A and B, where both peaks are fully separated (100% separation), so that the degree of overlap is zero (Ord, 1972; Evans et al., 2000).

3.2 Two Identical Square-Wave Profiles.

Figure 2(a–d) presents square-wave distribution profiles containing various degrees of overlap. The common practice in the literature is to calculate the percent overlap area and accordingly calculate the percent area separation for each peak consisting of a population of data points. The overlap in Figure 2(a–d) is represented by the shaded rectangular areas. The separation is one minus the ratio of overlap area to total area of each population (normally 1) as presented in eqs 1a, 1b, 2a, and 2b:

Percent overlap A = 100(overlap area/total area$_A$) **(1a)**
Percent overlap B = 100(overlap area/total area$_B$) **(1b)**

Percent separation A = 100(1 – overlap area/total area$_A$) **(2a)**
Percent separation B = 100(1 – overlap area/total area$_B$) **(2b)**



(a) **(b)** **(c)** **(d)**

Figure 2(a–d). Degrees of overlap for the A and B square-wave distributions of data having identical profiles. The shaded regions represent percent overlap of 0, 33, 50, and 100% for both square-wave data distributions in Figure 2(a–d), respectively.

Table 1 shows the values of the simple overlap area, percent overlap (eq 1), percent separation (eq 2), K-S test, and the ACD and LROC methods for the square-wave distribution profiles in Figure 2(a–d).Details of the K-S test and ACD and LROC methods are presented in Appendices A and B.

When two distributions of data are 100% separated, the values for the K-S test and ACD and LROC methods are 1.0, 0.5, and 2.0, respectively. When two separate distributions of data completely overlap (0% separation), the values for the K-S test and the ACD and LROC methods are 0.0, 0.0, and 1.41 ($\sqrt{2}$), respectively (shown in Appendices A and B and Fawcett, 2006; Hanley and McNeil, 1982; Wilcox, 1997; Press et al., 2007). For ease of plotting in a graphical format and comparison of all three statistical measures in a uniform manner, the plots are made

to start at 0.0. Therefore, because the K-S test and the ACD method start with 0.0 (complete overlap), the LROC method is modified to have a starting value with 0.0 denote 100% overlap by the simple subtraction of 1.41.

$$LROC_O = LROC - 1.41 \quad (3)$$

Table 1. Overlap Analyses of the Distributions of Data in Figure 2(a–d). The values of the overlap, percent overlap, percent separation, K-S test, and the ACD and $LROC_O$ methods for both square-wave distribution profiles are shown. Percent overlap and percent separation were calculated using eqs 1 and 2, respectively.

	Figure 2a	Figure 2b	Figure 2c	Figure 2d
Absolute Overlap Area	0	0.33	0.5	1.0
% Overlap A	0	33	50	100
% Overlap B	0	33	50	100
% Separation A	100	67	50	0
% Separation B	100	67	50	0
% Ave Separation (A+B)/2	100	67	50	0
K-S Test	1.0	0.67	0.5	0
ACD	0.5	0.45	0.38	0
LROC	2.0	1.8	1.71	1.41
$LROC_O$	0.59	0.39	0.30	0

The K-S test, the ACD, and $LROC_O$ methods overlap area values are plotted against the percent average separation (eq 2) as shown in Figure 3. The K-S test value indicates a strong linear correlation with percent average separation values ($r^2 = 1$). R-squared is the coefficient of determination (a measure of how well the model fits the data points).The $LROC_O$ value also provides a strong linear relationship with the percent average separation values ($r^2 = 1$). However, the ACD values show a strong quadratic relationship with the percent average separation values ($r^2 = 1$).

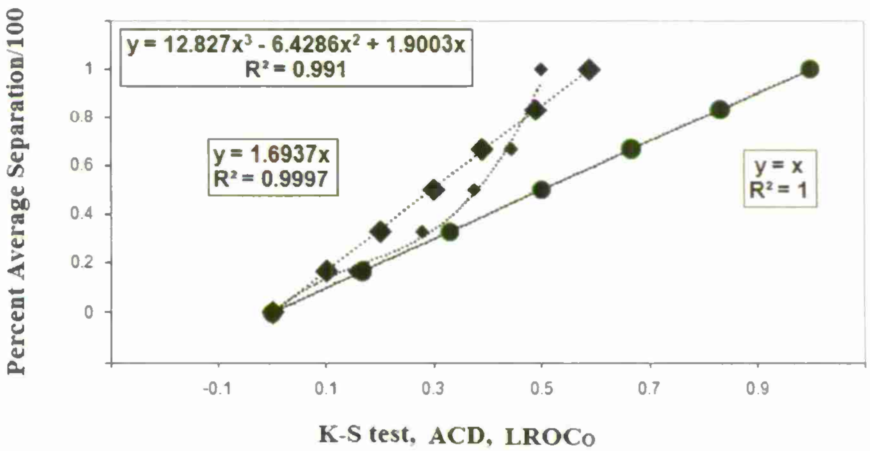


Figure 3. Relationships between the calculated percent average separation (using overlap area) vs. the K-S test (circles), ACD (small diamonds), and $LROC_O$ (big diamonds) values from

Table 1. The K-S test and $LROC_O$ method provide a strong linear correlation, whereas the ACD method shows a strong quadratic relationship.

Figure 2(a–d) and Table 1 illustrate that the results for all three methods are similar. Thus, all three methods identically track the simple separation statistics regardless of whether the shape is linear or curved. Therefore, for an analysis of two square-wave distributions having identical profiles, all methods yield identical degrees of separation.

3.3 Two Square-Wave Profiles with Different Peak Widths.

Attention is now drawn to the profile distribution of the two data sets in Figure 1b. The two populations have square-wave (rectangular) but differ in FWHH. In this example, the FWHH population for profile B is twice that of profile A. Figure 4(a–d) shows the square-wave distribution profiles with different FWHH values at various stages of overlap.

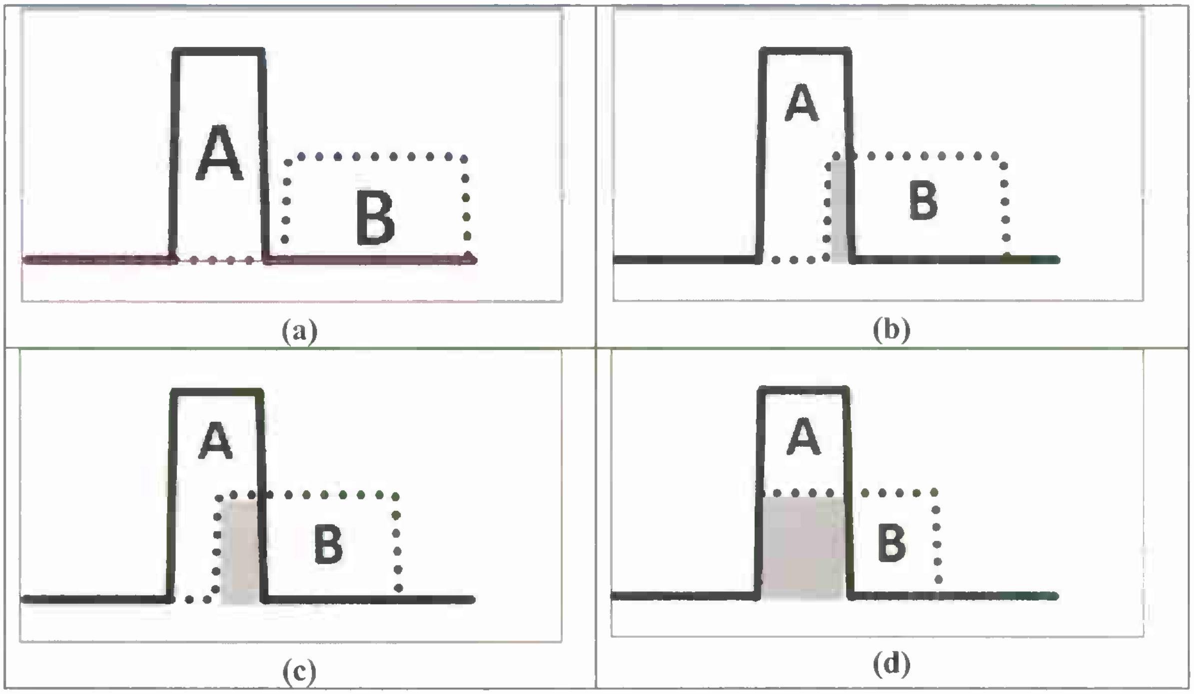


Figure 4(a–d). Degrees of overlap for two different data distribution profiles. Two populations of data are depicted as rectangular profiles but differ in FWHH. The B population FWHH is twice that of profile A. Different stages of distribution overlap are presented.

Unfortunately, the overlap area from eqs 1 and 2 cannot be used to calculate the degree of separation in the distribution profiles in Figure 4(b–d). Figure 4d shows that profile B is 50% fully separated and profile A is fully mixed with profile B. For any point inside the overlap region (gray area) in Figure 4(b–d), the probability of that point belonging to profile A (P_A) is 0.667 or 66.7%. In the gray area, the probability of that point belonging to profile A (P_A) is shown by eq 4 and the probability of that point belonging to profile B (P_B) is shown by eq 5.

$$P_A/(P_A + P_B) = 2/(2 + 1) \quad (4)$$

$$0.333 \text{ or } 33.3\% = P_B/(P_A + P_B) = 1/(2 + 1) \quad (5)$$

These are not relative values. On the contrary, the probability calculations are accurate and true assessments of the distribution of data points that resulted in the data profile. This information can be obtained for every set of experiments and it is a compact way of defining two data distributions. Therefore, probability distributions are essentially a reference database or a "lookup" distribution for the particular experiment. These are used as reference databases for actual unknown or blind experimental investigations.

3.4 Probability Functions as Reference Sources from Experimental Data.

It is clear that there is a need for a better approach to evaluate the degree of separation of both distributions. To accurately measure the degree of separation for each distribution with a 90% confidence or higher, the probability at each point on the x axis should be 0.9 or higher relative to all overlapped regions (gray area). In mathematical terms, the equation for calculating the degree of separation for each distribution (profiles A or B) with 90% confidence is derived as follows.

The probability density function, $P_A(x)$, for distribution A in Figure 4(a–d) that results in a 90% confidence ratio (CR) or higher should meet the following requirement:

$$CR_A(x) = P_A(x)/[P_A(x) + P_B(x)] \geq 0.9 \qquad \textbf{(6)}$$

The probability density function (Abramowitz, 1972; Ushakov, 2001) is merely the frequency rate (y-axis value) at a given number on the x axis that bisects profiles A and B. Thus, if the frequency rates (y-axis values) of profiles A and B are 2 and 10, respectively, at a given x-axis point of experimental value, then the $CR_A(x)$ for profile A equals 2/(2 + 10) or 16.7% confidence. Additionally, each x-axis value is examined for a CR value for profiles A and B so that the $CR_A(x)$ values are 0.90 or higher. This may or may not occur in the entire x-axis range of the data set.

The probability density function for distribution B or $P_B(x)$ with a 90% confidence or higher should meet the following requirement:

$$CR_B(x) = P_B(x)/[P_A(x) + P_B(x)] \geq 0.9 \qquad \textbf{(7)}$$

$P_A(x)$ and $P_B(x)$ are the original probability (frequency rate) values of distributions A and B, respectively, at any value on the x axis.

If at least one x value is common to both profile distributions (mixing areas or gray areas), the separated area with 90% confidence for each distribution will be calculated using eqs 8 and 9.

The free (separated or no-overlapped) area of distribution A with 90% confidence that is free of B points is

$$\mathbf{A_{90}} = {}_{-\infty}\int^{+\infty} F(x)\,dx \qquad \textbf{(8)}$$

where

$F(x) = P_A(x)$ if $CR(x)_A \geq 0.9$
$F(x) = 0$ if $CR(x)_A < 0.9$

The free (separated or nonoverlapped) area of distribution B with a 90% confidence that is free of A points is

$$\mathbf{B_{90}} = {}_{-\infty}\int^{+\infty} F(x)\,dx \qquad \textbf{(9)}$$

where

$F(x) = P_B(x)$ if $CR(x)_B \geq 0.9$
$F(x) = 0$ if $CR(x)_B < 0.9$

$\mathbf{A_{90}}$ and $\mathbf{B_{90}}$ equal 1.0 when both distributions do not overlap. This is the case for a complete separation of the two distributions.

The average percent separation with 90% confidence using eqs 8 and 9 is

$$\mathbf{AB_{90}} = 100(\mathbf{A_{90}} + \mathbf{B_{90}})/2 \qquad \textbf{(10)}$$

Probability distributions can be modeled from the raw data and used as a reference database, but the objective of the work herein was to find a method that resolves two distributions of data with only one or a few experimental determinations and without the need for constructing or modeling a probability distribution. In practice, to model a probability distribution we need at least 500–1000 measurements, but obtaining 500–1000 measurements for each distribution is neither practical nor possible.

3.5 Application of Probability Functions on Square-Wave Profiles.

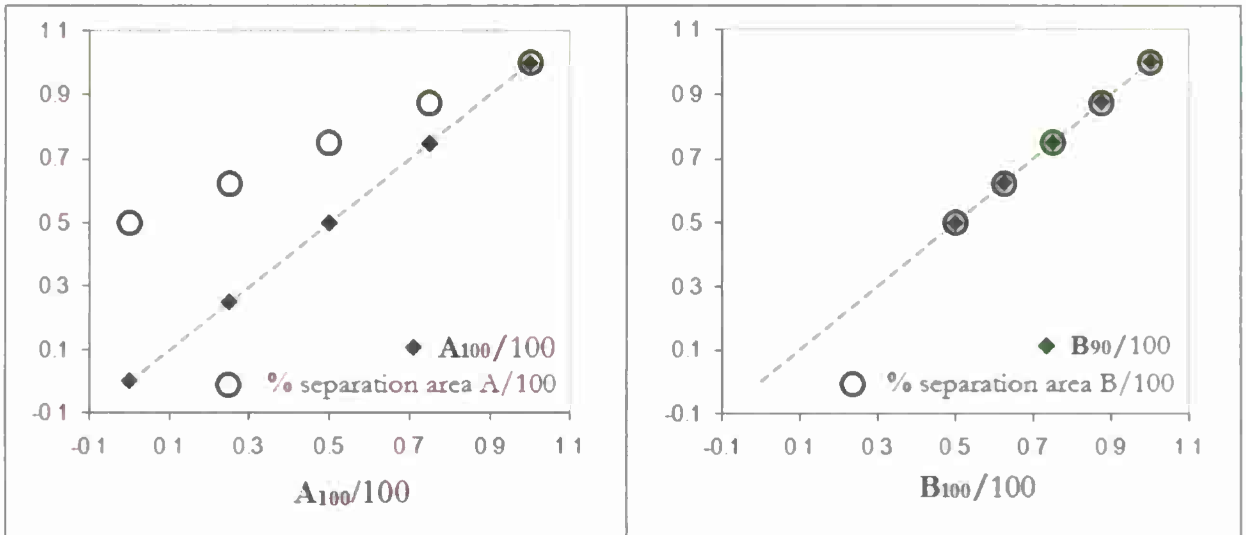
Table 2 lists various calculated values of the overlap area, percent overlap, percent separation using the overlap area, values of the $\mathbf{A_{90}}$ and $\mathbf{B_{90}}$ for distributions A and B, respectively, and the $\mathbf{AB_{90}}$ value for the Figure 2 data. The average percent separation of 100 $\mathbf{A_{90}}$ and 100 $\mathbf{B_{90}}$ values will be defined as $100(\mathbf{A_{90}} + \mathbf{B_{90}})/2 = 50(\mathbf{A_{90}} + \mathbf{B_{90}}) = \mathbf{AB_{90}}$. Table 2 also lists the calculated area for 100% separation confidence for distributions A ($\mathbf{A_{100}}$) and B ($\mathbf{B_{100}}$).

Table 2. Calculated Values of the Overlap Area, Percent Overlap, and Percent Separation Using the Overlap Area from Figure 2(a–d). Also listed are the $\mathbf{A_{90}}$ and $\mathbf{B_{90}}$ values for distributions A and B, respectively, and the $\mathbf{AB_{90}}$ values.

	Figure 4a	Figure 4b	Figure 4c	Figure 4d
Overlap Area	0	0.13	0.25	0.5
% Overlap A	0	13	25	50
% Overlap B	0	13	25	50
% Separation A	100	87	75	50
% Separation B	100	87	75	50
% Ave Separation (A+B)/2	100	87	75	50
$100\mathbf{A_{90}}$	100	75	50	0
$100\mathbf{B_{90}}$	100	87	75	50
$\mathbf{AB_{90}}$	100	81	62.5	25
$100\mathbf{A_{100}}$	100	75	50	0
$100\mathbf{B_{100}}$	100	87	75	50
$\mathbf{AB_{100}}$	100	81	62.5	25

Table 2 calculated (distribution A) values of percent separation using overlap area and $\mathbf{A_{90}}$ are plotted against the $\mathbf{A_{100}}$ values (Figure 5a). Table 2 calculated (distribution B) values of percent separation using overlap area, and $\mathbf{B_{90}}$ are plotted against the $\mathbf{B_{100}}$ values (Figure 5b). The $\mathbf{A_{100}}$ and $\mathbf{B_{100}}$ values are the true reference values for degree of separation for each profile.

A probability function such as $P_A(x)$ or $P_B(x)$ is plotted as a continuous function (algorithm). These values are not calculated as $\mathbf{A_{90}}$, $\mathbf{A_{100}}$, $\mathbf{B_{90}}$, $\mathbf{B_{100}}$, and $\mathbf{AB_{90}}$; rather, $P(x)$ functions are derived from experimental, replicate data and are hence plotted. Probability functions are unlike the simple overlap areas of A and B or ratio values such as $\mathbf{A_{90}}$, $\mathbf{A_{100}}$, $\mathbf{B_{90}}$, $\mathbf{B_{100}}$, and $\mathbf{AB_{90}}$, which are discrete calculated values.



a. (A distribution profile) **b.** (B distribution profile)

Figure 5(a,b). True reference value for degree of separation with 100% confidence separation (a) for distribution A ($\mathbf{A_{100}}$) and (b) for distribution B ($\mathbf{B_{100}}$) (*x* axis) vs. the (a) percent separation using overlap area A/100 (open circles, *y* axis) and the 90% confidence $\mathbf{A_{90}}$/100 values (closed circles) and (b) the percent separation B/100 (closed circles) and 90% confidence values $\mathbf{B_{90}}$/100 (closed circles) (*y* axis), respectively. The diagonal lines represent (a) $\mathbf{A_{100}}$ vs. $\mathbf{A_{100}}$ and (b) $\mathbf{B_{100}}$ vs. $\mathbf{B_{100}}$. Closed circles represent (a) $\mathbf{A_{100}}$/100 vs. $\mathbf{A_{90}}$/100 and (b) $\mathbf{B_{100}}$/100 vs. percent separation B/100 and $\mathbf{B_{90}}$/100.

It is clear that using the overlap area for measuring the degree of separation for distribution A profile overestimates (open circles) the percent separation.

3.6 Application of Probability Functions on Square-Wave Profiles with Different Peak Widths.

Table 3 shows the calculated values of $\mathbf{A_{90}}$ and $\mathbf{B_{90}}$ for distributions A and B, respectively, the average percent separation with 90% confidence ($\mathrm{AB_{90}}$), and the K-S test, ACD, and $\mathrm{LROC_O}$ values for the square-wave distribution profiles in Figure 4(a–d). The average percent separation of 100$\mathbf{A_{90}}$ and 100$\mathbf{B_{90}}$ is defined as $100(\mathbf{A_{90}} + \mathbf{B_{90}})/2 = 50(\mathbf{A_{90}} + \mathbf{B_{90}}) = \mathbf{AB_{90}}$ as shown by eq 10.

Figure 6 shows the relationship between the $\mathbf{AB_{90}}$ and the K-S test, ACD, and $\mathrm{LROC_O}$ values. The calculated values in Table 3 of the K-S test and the ACD and $\mathrm{LROC_O}$ methods are plotted against the $\mathbf{AB_{90}}$ values in Figure 6. The K-S test shows a strong quadratic correlation with the $\mathbf{AB_{90}}$ values with $r^2 = 1$. The $\mathrm{LROC_O}$ values provide a strong linear correlation with the $\mathbf{AB_{90}}$ values and $r^2 = 1$. The ACD shows a strong third-order relationship with the $\mathbf{AB_{90}}$ values and $r^2 = 0.99$. For the distributions in Figure 4(a–d), all measures were virtually identical in information content because relatively simple data profiles were used (square-wave distribution).

Table 3. Overlap Analyses of the Two Distributions of Data in Figure 4(a–d). The calculated values of $\mathbf{A_{90}}$, $\mathbf{B_{90}}$, $\mathbf{AB_{90}}$, and the K-S test, ACD, and $LROC_O$ methods are shown.

	Figure 4a	Figure 4b	Figure 4c	Figure 4d
$100\mathbf{A_{90}}$	100	72.5	47.5	0
$100\mathbf{B_{90}}$	100	86.3	74	50
$\mathbf{AB_{90}}$	100	79.4	60.6	25
K-S test	1.0	0.875	0.75	0.5
ACD	0.5	0.48	0.44	0.25
LROC	2.0	1.91	1.8	1.62
$LROC_O$	0.59	0.5	0.39	0.21

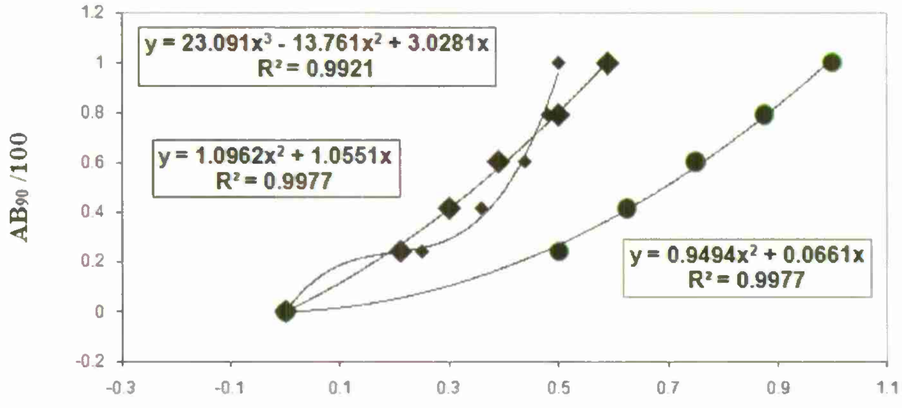


Figure 6. Relationships between $\mathbf{AB_{90}}$ vs. the K-S test (circles), ACD (small diamonds), and $LROC_O$ (large diamonds) values for Figure 4(a–d). The $LROC_O$ method sows a strong linear correlation, whereas the K-S test shows a strong quadratic relationship.

3.7 Application of Probability Functions on Triangular Data Profiles.

The triangular distributions in Figure 1c closely resemble a binomial distribution. Figure 7(a–d) shows triangular-distribution profiles with the same FWHH at various overlap situations. Note that a gray area of overlap is not outlined in Figure 7(a–d). Instead, the measure is for specific x-axis points, one at a time, to be analyzed by eqs 8 and 9 for a probability of an x-axis point attaining a CR value of 0.9 or higher for either distribution A or B.

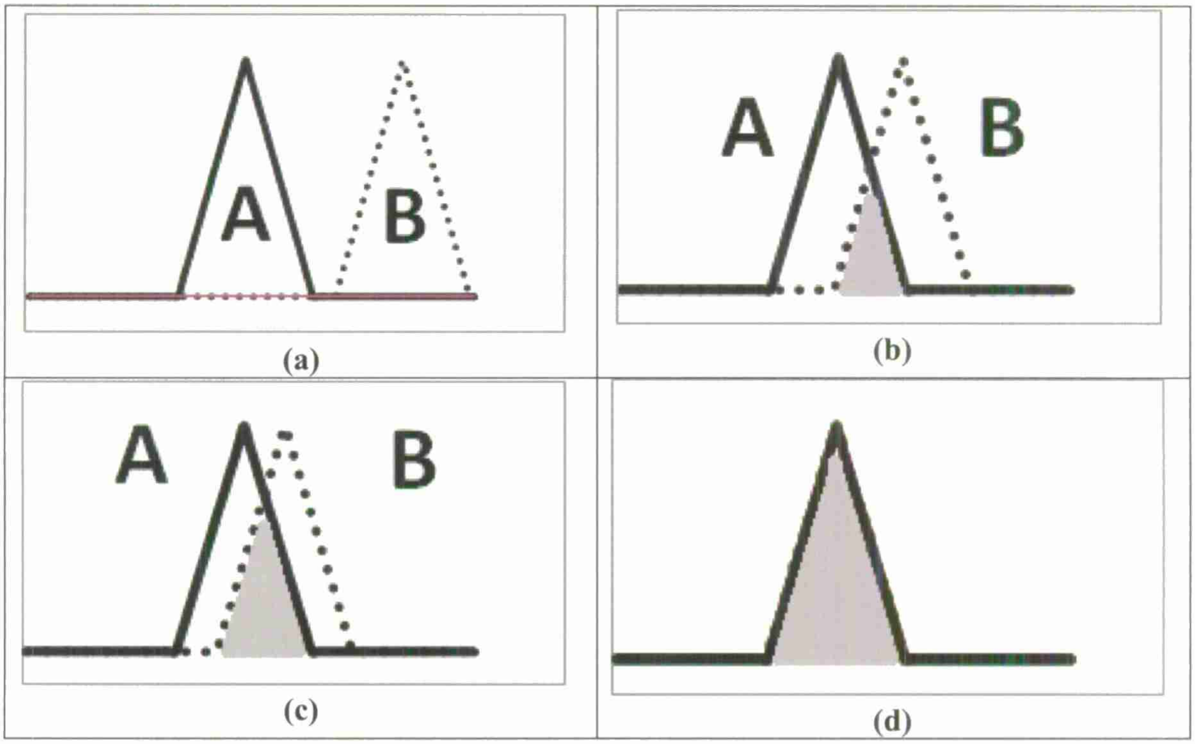


Figure 7(a–d). Degrees of overlap for two data distributions with the same triangular profile. The two triangular distributions of data have the same FWHH. The distributions closely resemble the binomial distribution.

Table 4 shows the calculated $\mathbf{A_{90}}$, $\mathbf{B_{90}}$, $\mathbf{AB_{90}}$, K-S test, ACD, and $LROC_O$ values for the distributions in Figure 7(a–d).

Table 4. Overlap Analyses of the Distributions of Data in Figure 7. The calculated values for $\mathbf{A_{90}}$, $\mathbf{B_{90}}$, $\mathbf{AB_{90}}$, the K-S test, and the ACD and $LROC_O$ methods are shown.

	Figure 7a	Figure 7b	Figure 7c	Figure 7d
$100\mathbf{A_{90}}$	100	50	22	0
$100\mathbf{B_{90}}$	100	50	22	0
$\mathbf{AB_{90}}$	100	50	22	0
K-S test	1.0	0.72	0.51	0
ACD	0.5	0.45	0.34	0
LROC	2.0	1.79	1.64	1.41
$LROC_O$	0.59	0.38	0.23	0

The Table 4, calculated values of the K-S test and the ACD and $LROC_O$ methods are plotted against the $\mathbf{AB_{90}}$ values (Figure 8). The K-S test and $LROC_O$ values yield strong quadratic correlations with the $\mathbf{AB_{90}}$ values ($r^2 = 1$). The ACD values show a strong third-order relationship with the $\mathbf{AB_{90}}$ values ($r^2 = 0.98$).

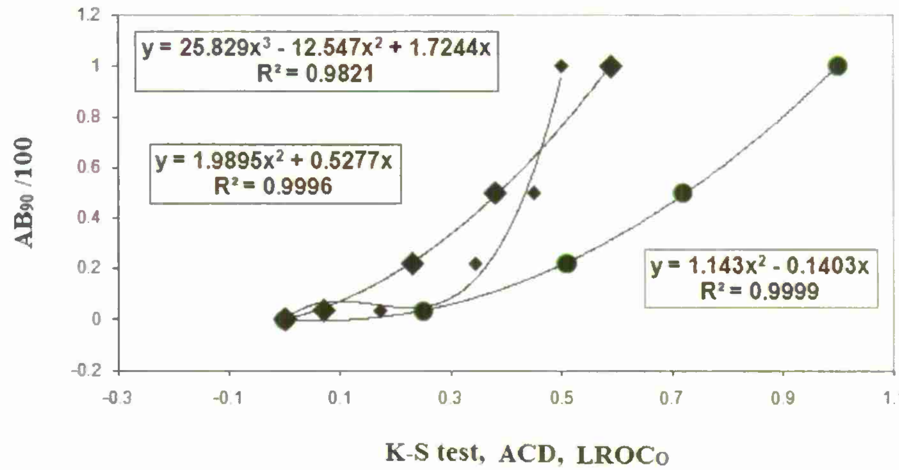


Figure 8. Relationships between the $\mathbf{AB_{90}}$ vs. the K-S test (small diamonds), ACD (circles), and $LROC_O$ (large diamonds) values from the Table 4 data. The K-S test and the $LROC_O$ method provide a strong quadratic relationship.

3.8 Application of Probability Functions on Two Triangular Profiles with Different Peak Widths.

Now a more complex, yet practical, distribution such as that shown in Figure 1d is presented. Figure 9(a–d) shows triangular-distribution profiles with different FWHH at various degrees of overlap.

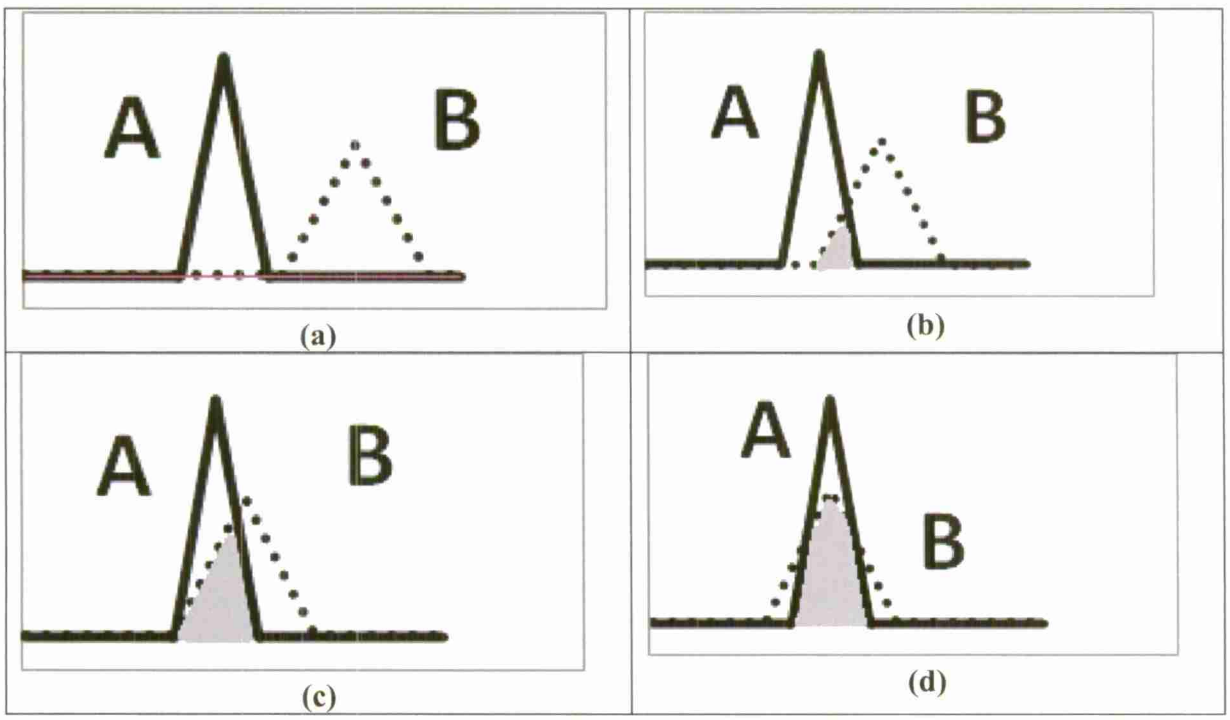


Figure 9(a–d). Degrees of overlap for two different distributions of data. Two populations display triangular-distribution profiles with different FWHH, at different degrees of overlap.

Table 5 shows the calculated values from the $\mathbf{A_{90}}$, $\mathbf{B_{90}}$, $\mathbf{AB_{90}}$, K-S test, ACD, and $LROC_O$ analyses for the Figure 9(a–d) data.

Table 5. Overlap Analyses of the Two Distributions of Data in Figure 9(a–d) for the $\mathbf{A_{90}}$, $\mathbf{B_{90}}$, $\mathbf{AB_{90}}$, K-S Test, ACD, and $LROC_O$ Values.

	Figure 9a	Figure 9b	Figure 9c	Figure 9d
100$\mathbf{A_{90}}$	100	37.5	0	0
100$\mathbf{B_{90}}$	100	65	32	16
$\mathbf{AB_{90}}$	100	51.3	16	8
K-S test	1.0	0.74	0.47	0.125
ACD	0.5	0.44	0.29	0.04
LROC	2.0	1.8	1.6	1.49
$LROC_O$	0.59	0.39	0.18	0.08

The calculated values from the K-S test and the ACD and $LROC_O$ methods are plotted against the $\mathbf{AB_{90}}$ values as shown in Figure 10. The K-S test and $LROC_O$ method display a strong quadratic correlation with the $\mathbf{AB_{90}}$ values (r^2 = 0.99 and 1, respectively). The ACD values yield a strong third-order relationship with the $\mathbf{AB_{90}}$ values (r^2 = 0.99).

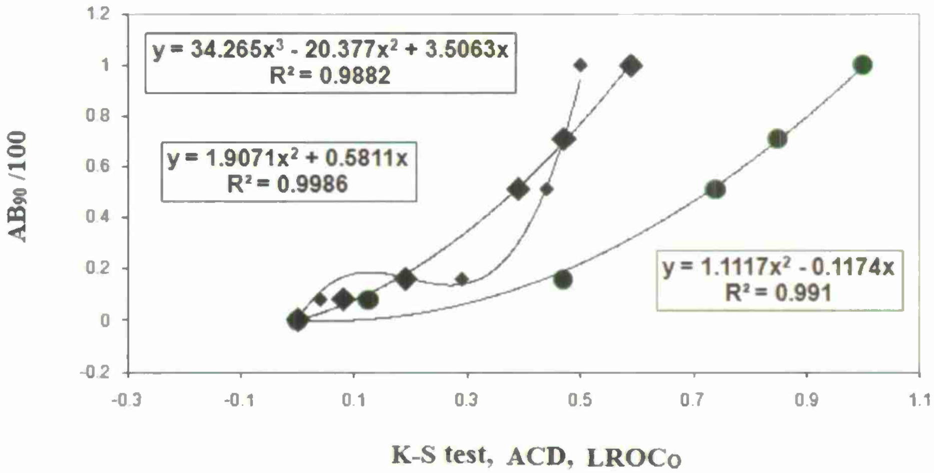


Figure 10. $\mathbf{AB_{90}}$ values plotted against the K-S test (circles), ACD (small diamonds), and $LROC_O$ (large diamonds) values from Table 5. The $LROC_O$ method and K-S test yield a strong quadratic relationship.

3.9 Application of Probability Functions on Skewed Profiles.

An even more difficult distribution profile is presented in Figure 1e. Both data populations are positive-skewed distribution (PSD) profiles with different FWHHs. Figure 11(a–d) shows the PSD distribution profiles with different FWHHs at various overlap stages.

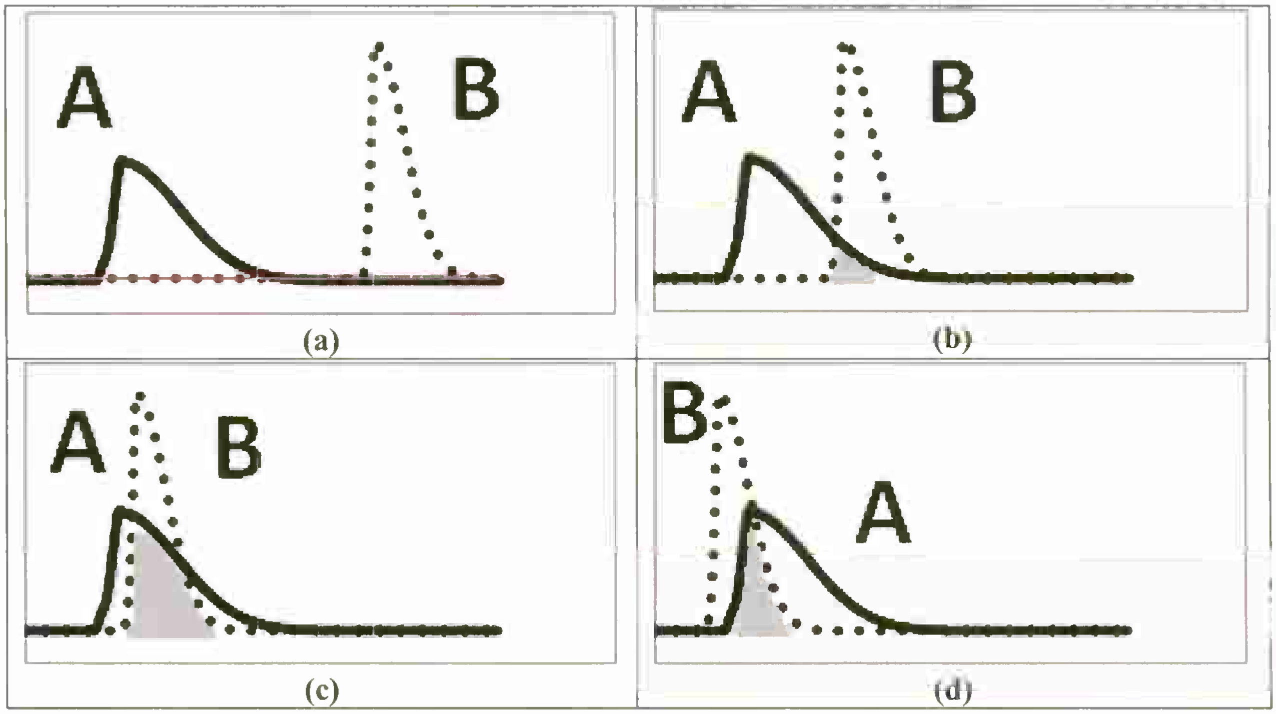


Figure 11(a–d). Degrees of overlap for two PSDs.

Table 6 shows the calculated values of the $\mathbf{A_{90}}$, $\mathbf{B_{90}}$, $\mathbf{AB_{90}}$, K-S test, ACD, and $LROC_O$ analyses for the PSD distribution profiles in Figure 11(a–d).

Table 6. Overlap Analyses of Two Distributions of Data in Figure 11(a–d). $\mathbf{A_{90}}$, $\mathbf{B_{90}}$, $\mathbf{AB_{90}}$, K-S test, ACD, and $LROC_O$ values are shown.

	Figure 11a	Figure 11b	Figure 11c	Figure 11d
100$\mathbf{A_{90}}$	100	86	24.5	43
100$\mathbf{B_{90}}$	100	69	0	40
$\mathbf{AB_{90}}$	100	77.5	12.3	41.5
K-S test	1.0	0.87	0.25	0.61
ACD	0.5	0.44	0.07	0.41
LROC	2.0	1.88	1.55	1.72
$LROC_O$	0.59	0.47	0.14	0.31

The calculated values for the K-S test and the ACD and $LROC_O$ methods are plotted against the $\mathbf{AB_{90}}$ values as shown in Figure 12. The K-S test and $LROC_O$ method yield strong quadratic correlations with the $\mathbf{AB_{90}}$ values (r^2 = 0.96 and 0.98, respectively). The ACD values show a strong third-order relationship with the $\mathbf{AB_{90}}$ values (r^2 = 0.97).

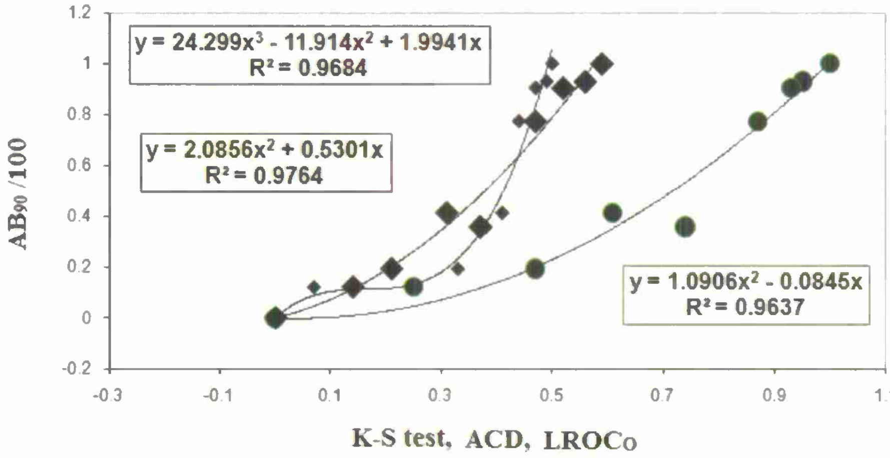


Figure 12. Relationships between the **AB$_{90}$** and the K-S test (circles), ACD (small diamonds), and LROC$_O$ (large diamonds) values in Table 6. The K-S test and LROC$_O$ method show strong quadratic relationships.

3.10 Application of Probability Functions on Triangular Profiles with a Bimodal Distribution.

The data distributions as shown in Figure 1f are examined with the separation algorithms. The distribution A population is a triangular-distribution profile (similar to a binomial distribution), whereas the distribution B population resembles a bimodal, negatively skewed distribution profile. Figure 13(a–d) shows both profiles at various overlap stages.

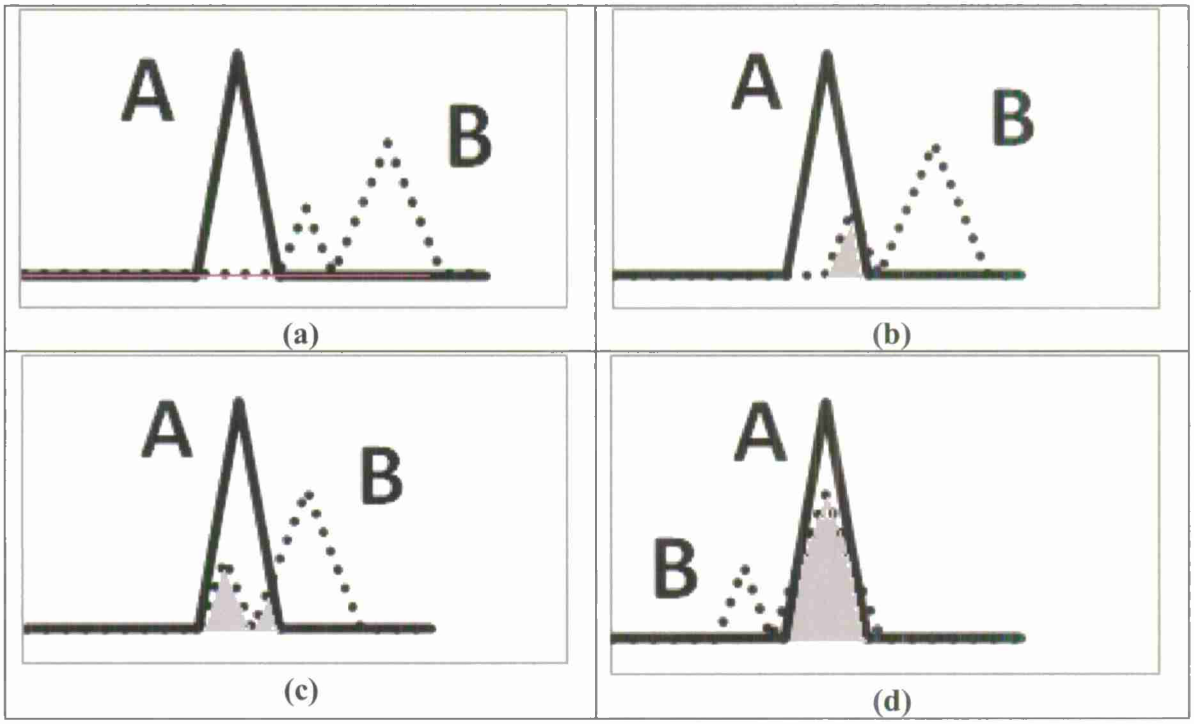


Figure 13(a–d). Degrees of overlap for two different distributions of data where profile A is a triangular distribution and profile B is a bimodal distribution.

Table 7 shows the calculated values of the $\mathbf{A_{90}}$, $\mathbf{B_{90}}$, $\mathbf{AB_{90}}$, K-S, ACD, and $LROC_O$ analyses from Figure 13(a–d).

Table 7. Overlap Analyses of the Distributions of Data in Figure 13(a–d). Calculated values for the $\mathbf{A_{90}}$, $\mathbf{B_{90}}$, $\mathbf{AB_{90}}$, K-S, ACD, and $LROC_O$ analyses are shown.

	Figure 13a	Figure 13b	Figure 13c	Figure 13d
$100\mathbf{A_{90}}$	100.0	62.5	20.0	0
$100\mathbf{B_{90}}$	100.0	82.0	70.0	26.5
$\mathbf{AB_{90}}$	100.0	72.3	45.0	13.3
K-S test	1.0	0.83	0.73	0.25
ACD	0.5	0.48	0.35	0.11
LROC	2.0	1.87	1.78	1.51
$LROC_O$	0.59	0.46	0.37	0.10

Calculated values for the K-S test and the ACD and $LROC_O$ methods are plotted against $\mathbf{AB_{90}}$ as shown in Figure 14. The K-S test and $LROC_O$ values show strong quadratic correlations with the $\mathbf{AB_{90}}$ (r^2 = 0.97 and 0.99, respectively). The ACD values show a strong third-order relationship with the $\mathbf{AB_{90}}$ values (r^2 = 0.97).

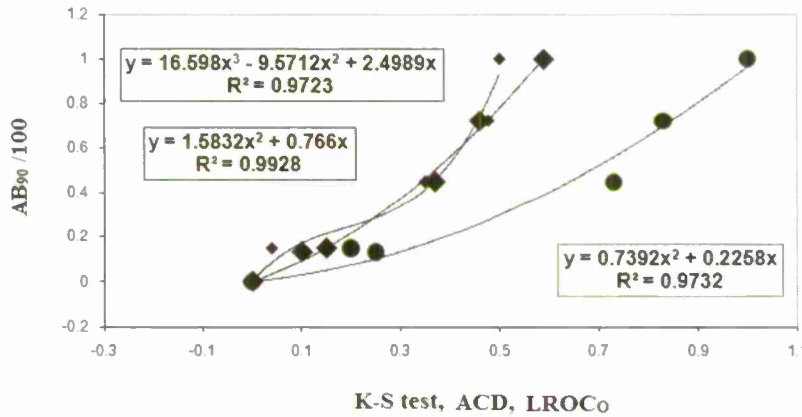


Figure 14. Relationships between the $\mathbf{AB_{90}}$ vs. the K-S test (circles), ACD (small diamonds), and $LROC_O$ (large diamonds) values in Table 7. The K-S test and $LROC_O$ method show a strong quadratic relationship.

3.11 Other Distribution Profiles.

Other complex distribution profiles were evaluated as shown in Figure 1(e–h). However, the objective of this report was to determine which of the three (K-S test, ACD, or $LROC_O$) approaches can predict more accurately the degree of separation between any two experimental populations of data regardless of the shape complexity of the distribution profiles.

All values from the K-S test and the ACD and $LROC_O$ methods calculated from Tables 1–6 and the equivalent analyses (data not shown) for Figure 1(g–i) at different degrees of overlap are plotted against the corresponding $\mathbf{AB_{90}}$ values and are shown in Figure 15. The

LROC$_O$ method clearly yields the best prediction comparison to the reference **AB$_{90}$** as opposed to the others, with a strong quadratic correlation ($r^2 = 0.99$). The next in line is the K-S test, which shows a strong quadratic relationship with the **AB$_{90}$** values ($r^2 = 0.92$). The ACD method shows a strong third-order relationship with the **AB$_{90}$** values ($r^2 = 0.90$). If the values of the K-S test and the ACD and LROC$_O$ methods from Figure 1(h,i) are removed from Figure 15, the values of r^2 become 0.98, 0.95, and 0.96 for the K-S test, ACD, and LROC$_O$ analyses, respectively, as shown in Figure 16.

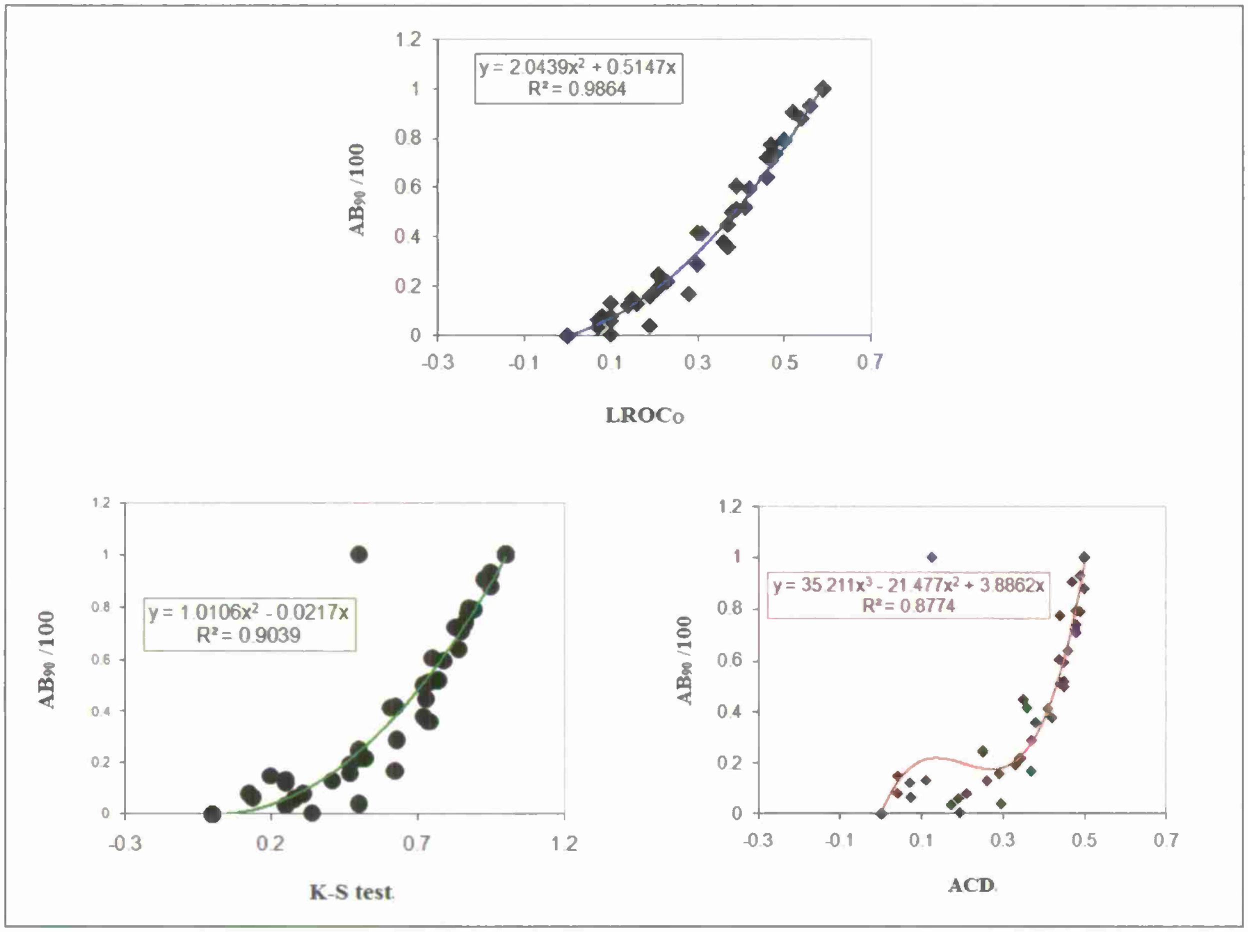


Figure 15. Plotted values from the K-S test, ACD, and LROC$_O$ analyses from Figure 1(a–i) at various overlap stages plotted against the **AB$_{90}$** values.

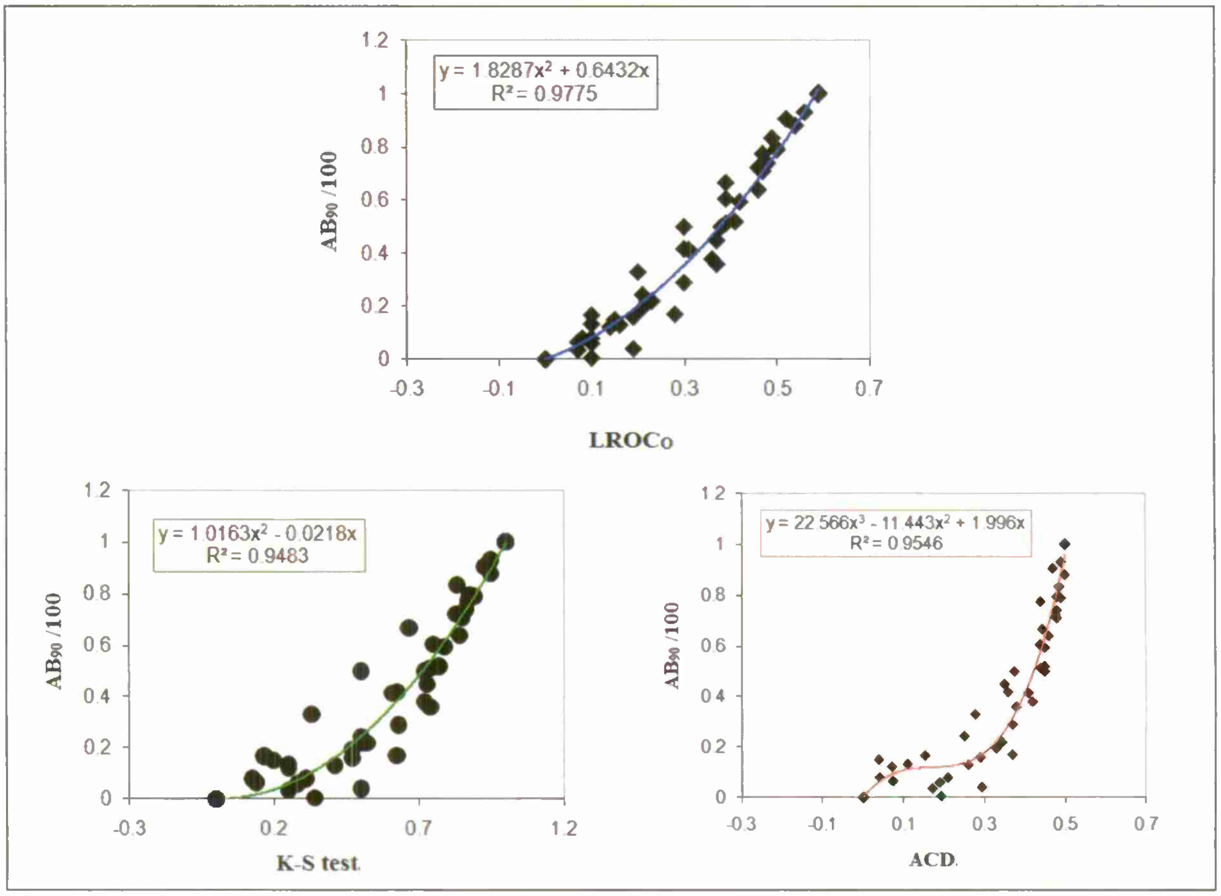


Figure 16. Figure 14 results with the data from Figure 1(h,i) removed.

The degree of separation is strongly dependent on the confidence limit set in the overlap (mixing) region between two distributions. Equations 11a–11c present a more accurate nonlinear fitting model than the third-order or second-order polynomial equations shown in Figures 15 and 16. The following nonlinear fitting model is used to calculate the average percent separation with respect to the ACD values between any two population distributions at 90% confidence. The nonlinear fitting models for the K-S test and LROC were generated (not shown here) similar to the ACD (eq 11).

$$\%\text{average separation @ 90\% confidence} = \text{function of ACD} \quad \textbf{(11a)}$$

$$\%\text{seperation}_{90} = \mathbf{AB_{90}} = f(\text{ACD}) \quad \textbf{(11b)}$$

$$y = 26.91\,x^3 - 25.51\,x^2 + 8.04x \qquad r^2 = 0.88 \quad \textbf{(11c)}$$

where

$$y = 1 - \mathbf{AB_{90}}/100 \qquad \text{and} \qquad x = 0.5 - \text{ACD}$$

4. CONCLUSIONS

Population profiles, varying from simple to complex distributions, were made to explore the capability of different methods to calculate the degree of separation of two distributions (two peaks or two populations). The most accurate measure for the degree of separation between any two distributions is derived by using basic statistics (i.e., eqs 8 and 9 at different confidence levels) where the probability density functions (A and B profiles) are known. If probability density functions are not available for use in judging actual experimental points, other statistical methods of separation analysis such as the K-S test and the ACD and $LROC_O$ methods must be used. The commonly used overlapped area between two distributions has proven to be inaccurate for measuring the degree of separation between two distributions (populations), even if the probability density functions are known. Equations 8 and 9 require the continuous probability functions of two distributions (A and B). In practice, and especially with a limited number of measurements (points), eqs 8 and 9 cannot be used.

Alternative methods (algorithms) were investigated to identify an optimal degree of separation between two distributions. The next best measure for the average degree of separation between two distributions, even for complex distributions, is the $LROC_O$ method. The advantage of the $LROC_O$ and ACD methods over the K-S test is that they offer additional statistical information (sensitivity, selectivity, and false alarm rate) in addition to a single nonparametric value (average percent separation).

Blank

LITERATURE CITED

Evans, M.; Hastings, N.; Peacock, B. Probability Density Function and Probability Function. *Statistical Distributions*; 3rd ed.; John Wiley & Sons: New York, 2000; 9–11.

Fawcett, T. An Introduction to ROC Analysis. *Pattern Recognition Letters* **2006**, *27*(8), 861–874.

Hanley J.A.; McNeil B.J. The Meaning and Use of the Area under the Receiver Operating Characteristic (ROC) Curve. *Radiology* **1982**, *143*(1), 29–36.

Metz, C.E. Basic Principles of ROC Analysis. *Seminars in Nuclear Medicine* **1978**, *8*, 283–298.

Ord, J.K. *Families of Frequency Distributions (Griffin's Statistical Monographs, No 30)*; Lubrecht & Cramer Ltd: New York, 1972.

Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P. *Numerical Recipes: The Art of Scientific Computing*; 3rd ed.; Cambridge University Press: New York, 2007.

Probability Functions. In *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*; Abramowitz, M., Stegun, I.A., Eds.; 9th printing; Dover Publications: New York, 1972; 925–964.

Ushakov, N.G. Density of a Probability Distribution. In *Encyclopaedia of Mathematics*; Hazewinkel, M., Ed.; Springer Publishing Company: New York, 2001.

Wilcox, R.R. Some Practical Reasons for Reconsidering the Kolmogorov-Smirnov Test. *British Journal of Mathematical and Statistical Psychology* **1997**, *50*(1), 71–78.

GLOSSARY

$\mathbf{A}_{90}$	free (separated or nonoverlapped) area of distribution A with a 90% confidence that is free of B points
$\mathbf{A}_{100}$	free (separated or nonoverlapped) area of distribution A with a 100% confidence that is free of B points
$\mathbf{AB}_{90}$	$100(\mathbf{A}_{90} + \mathbf{B}_{90})/2$, the average percent separation of profiles A and B with a 90% confidence.
$\mathbf{AB}_{100}$	$100(\mathbf{A}_{100} + \mathbf{B}_{100})/2$, the average percent separation of profiles A and B with a 100% confidcnce
ACD	ROC curve referenced to the diagonal line
AUC	area under the curve
$\mathbf{B}_{90}$	free (separated or nonoverlapped) area of distribution B with a 90% confidence that is free of A points
$\mathbf{B}_{100}$	free (separated or nonoverlapped) area of distribution B with a 100% confidence that is free of A points
$\mathrm{CR}_{\mathbf{A}}(x)$	confidence ratio for a point A belonging in distribution A in an overlap area
$\mathrm{CR}_{\mathbf{B}}(x)$	confidence ratio for a point B belonging in distribution B in an overlap area
D	difference of two cumulative distribution functions
FN	false negative
FP	false positive
FWHH	full width at half height
K-S test	Kolmogorov-Smirnov test
LROC	length of the ROC curve
$\mathrm{LROC}_{\mathrm{O}}$	= LROC – 1.41
MVA	multivariate data analysis
ROC	receiver operating characteristic
$P_{\mathrm{A}}(x)$	probability of a point belonging to distribution A of points
$P_{\mathrm{B}}(x)$	probability of a point belonging to distribution B of points
PSD	positive-skewed distribution
$\mathrm{S}_{\mathbf{N1}}(x)$	cumulative (additive) distribution function
TN	true negative
TP	true positive

APPENDIX A
KOLMOGOROV-SMIRNOV TEST

The K-S test is a goodness-of-fit test used to assess the uniformity of a set of data distributions. The K-S test is applied to determine whether two data sets (*x* and *y*) differ significantly. The K-S test has an advantage in that no assumptions are made about the distribution of data. Technically speaking, it is a nonparametric algorithm that is independent of distribution tendencies. In statistics, the K-S test is the accepted test for measuring differences between continuous data sets (unbinned data distributions) that are a function of a single variable. The difference is defined as the maximum (linear) value of the absolute difference between two cumulative distribution functions. Thus, for comparing two different cumulative distribution functions $S_{N1}(x)$ and $S_{N2}(x)$, the K-S statistic is

$$D = \max_{-\infty < x < \infty} |S_{N1}(x) - S_{N2}(x)|$$

Figure A1 shows an example of the *x*-variable population (dotted line) and the *y*-variable population (solid line) in a cumulative fraction plot. The K-S test shows the variation or difference (D) between the two variables (*x* and *y*) and the measure of how they differ (D). Figure A2 shows that the *x*-variable population (dotted line) has more variation than the *y*-variable population (solid line), even though they have a similar median.

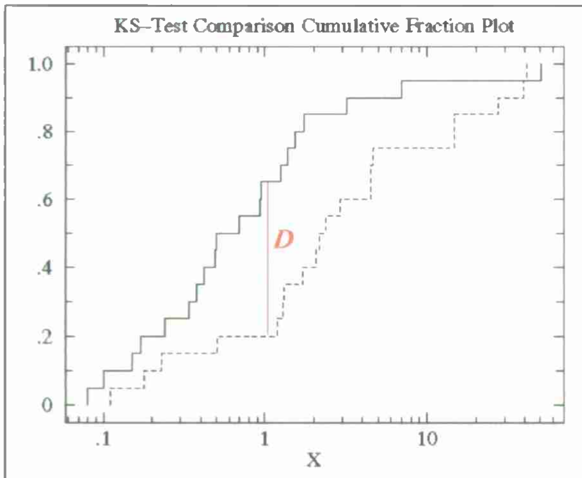


Figure A1. An example of an *x*-variable population (dotted line) and a *y*-variable population (solid line) in a cumulative fraction plot. The K-S test shows the variation (D) between the two variables (*x*,*y*) and the measure of how they differ (D).

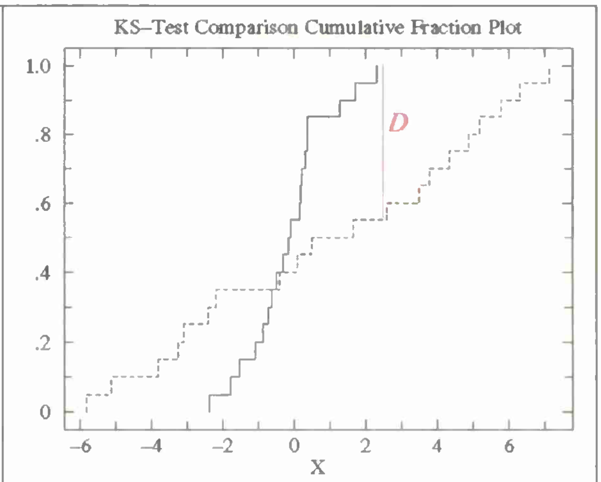
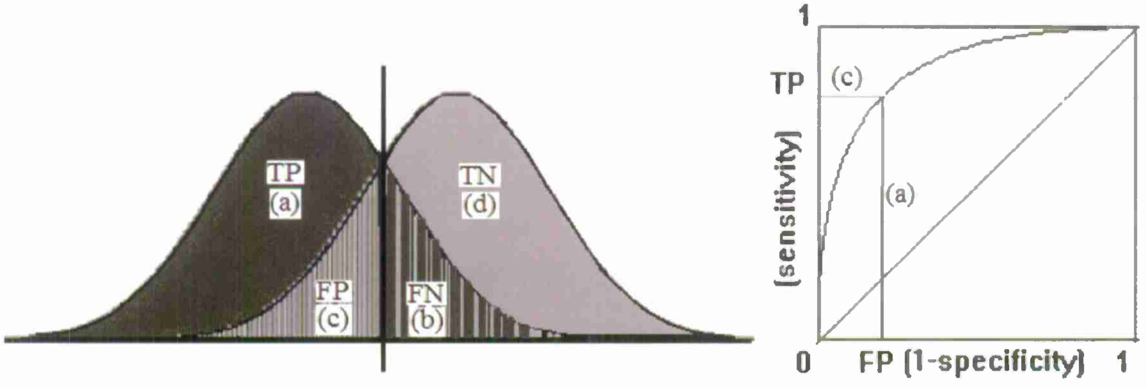


Figure A2. A plot of two different data distributions. The plot shows that the *x*-variable population (dotted line) has more variation than the *y*-variable population (solid line), even though they have similar medians.

Blank

APPENDIX B
THE AREA BETWEEN THE ROC CURVE AND
THE DIAGONAL LINE ACD AND LENGTH OF THE ROC CURVE

A brief discussion of the receiver operating characteristic (ROC) curve methodology is presented. Two distributions in Figure B1 are partially overlapped. For every possible cut-off point (a vertical line at any *x* value through the Gaussian curves in Figure B1) that is chosen to discriminate between the two populations, there will be some cases where the dark-gray value is correctly classified as a dark-gray point (TP = true positive fraction, a); and some cases where the dark-gray will be incorrectly classified as a light-gray point (FN = false negative fraction, b). On the other hand, some cases with the light-gray will be correctly classified as light-gray points (TN = true negative fraction, d), but some cases will occur where a light-gray point will be incorrectly classified as a dark-gray point (FP = false positive fraction, c). The ROC curve can be created by simply plotting TP (a) versus FP (c).



1st Distribution Example Set of Data	2nd Distribution Example Set of Data
Dark-Gray	Light-Gray
Diseased	Healthy
Positive	Negative
Present	Absent
Group-A	Group-B
Background	Sample

Figure B1. Two distributions of related data that display a partial overlap of results. For every possible cut-off point of an *x*-value selected to discriminate between two regions, some dark-gray cases will be correctly (positively) classified as dark-gray points (TP); and in some cases, a light-gray data point will be incorrectly classified as a dark-gray point (FP). The ROC curve is created by simply plotting TP (a) vs. FP (c).

Table B lists all parameters related to the ROC curve generated from Figure B1. The area under the ROC curve (AUC, all the way to the baseline or *x*-axis) can be simply calculated by an extended trapezoidal rule.* Figure B2 shows how the area is calculated between the ROC curve and the diagonal line (ACD). The diagonal line (red line shown in Figure B2) is a line of no separation (discrimination) between the two groups.

Table B. Major Variables (TP, FP, TN, FN) of ROC curve statistics (Metz, 1978; Hanley and McNeil, 1982)

	Dark-Gray area	**#**	**Light-Gray area**	**#**	**Total #**
	TP or true dark-gray	a	FP or false dark-gray	c	a + c
	FN or false light-gray	b	TN or true light-gray	d	b + d
Total		a + b		c + d	

Sensitivity	a/(a + b)	Specificity	d/(c + d)
Positive Likelihood ratio	Sensitivity/ (1 – specificity)	Negative Likelihood ratio	(1 – Sensitivity)/ (Specificity)
Positive Predictive value	a/(a + c)	Negative Predictive value	d/(b + d)

- Sensitivity: probability that a test result will be positive when the dark-gray is present (TP rate, expressed as a percentage) = a/(a + b)
- Specificity: probability that a test result will be negative when the dark-gray is not present (TN rate, expressed as a percentage) = d/(c + d)
- Positive likelihood ratio: ratio between the probability of a positive test result given the presence of the dark-gray and the probability of a positive test result given the absence of the dark-gray, i.e., = TP rate/FP rate = sensitivity/(1 – specificity)
- Negative likelihood ratio: ratio between the probability of a negative test result given the presence of the dark-gray and the probability of a negative test result given the absence of the dark-gray, i.e., = FN rate/TN rate = (1 – sensitivity)/specificity
- Positive predictive value: probability that the dark-gray is present when the test is positive (expressed as a percentage) = a/(a + c)
- Negative predictive value: probability that the dark-gray is not present when the test is negative (expressed as a percentage) = d/(b + d)
- Calculated area under the curve referenced to the diagonal line (ACD)

* Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P. *Numerical Recipes: The Art of Scientific Computing*; Third Edition; Cambridge University Press: New York, 2007; p 1235.

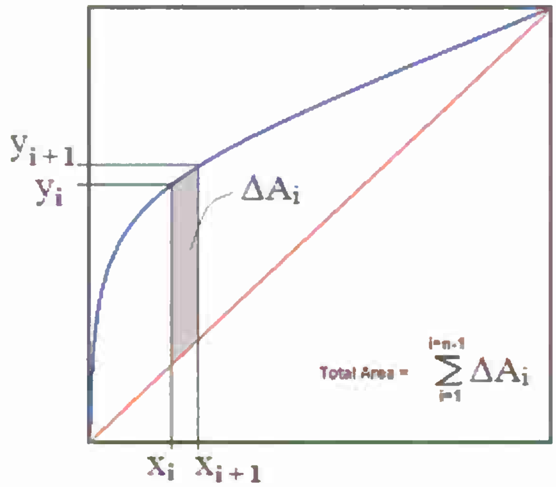


Figure B2. Integration algorithm showing how the ACD is calculated from the ROC curve. The diagonal line (red line) is a line of no separation (discrimination) between the two groups.

The ACD is calculated by summing all small grey areas (ΔA_i) depicted in Figure B2 between two adjacent points as shown mathematically by eq B1. The length of the ROC curve (LROC) is calculated by simply summing all the small incremental lengths between any two adjacent points on the ROC curve as shown mathematically by eq B2.

$$ACD = \sum (x_{i+1} - x_i)[(y_{i+1} + y_i) - (x_{i+1} + x_i)]/2 \qquad \text{(B1)}$$

$$LROC = \sum \sqrt{}(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 \qquad \text{(B2)}$$

EDGEWOOD CHEMICAL BIOLOGICAL CENTER

A U.S. Army RDECOM Laboratory